IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 November 2022, accepted 14 December 2022, date of publication 19 December 2022,
date of current version 27 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3230592

RESEARCH ARTICLE

ETFPOS-IDF: A Novel Term Weighting Scheme for Examination Question Classification Based on Bloom's Taxonomy

MOHAMMED OSMAN GANI[1], RAMESH KUMAR AYYASAMY[2], (Senior Member, IEEE), SAADAT M. ALHASHMI[3], ANBUSELVAN SANGODIAH[4], AND YONG TIEN FUI[2]

[1]Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar 31900, Malaysia
[2]Department of Information Systems, Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar 31900, Malaysia
[3]Department of Information Systems, University of Sharjah, Sharjah, United Arab Emirates
[4]School of Computing, Faculty of Computing and Engineering, Quest International University, Ipoh, Perak 30250, Malaysia

Corresponding authors: Ramesh Kumar Ayyasamy (rameshkumar@utar.edu.my) and Saadat M. Alhashmi (salhashmi@sharjah.ac.ae)

This work was supported by the Universiti Tunku Abdul Rahman (UTAR) Research Fund (UTARRF) under Grant IPSR/RMC/UTARRF/2020-C2/A01.

ABSTRACT Numerous earlier studies focused on the term weighting scheme to increase examination question classification accuracy based on Bloom's Taxonomy (BT). While determining the cognitive level of the examination question, all the terms present in the question are not equally significant. Verbs are the most important parts of speech while assigning weights to the terms. However, two types of verbs may be present in the questions: BT and supporting. BT verbs have a higher impact on determining the cognitive level of a question than supporting verbs. Nevertheless, the proposed schemes of past studies assigned equal weight to both types of verbs. Therefore, this study aims to introduce the term weighting scheme ETFPOS-IDF, which assigns BT a higher weight than supporting verbs. The BT verbs were identified based on their position in the questions. Three datasets and three classifiers: Support Vector Machine, Artificial Neural Network, and Random Forest, were used in this study. Two evaluation metrics: accuracy and F1 score, were used to evaluate the performance of the proposed model. The experiment results showed that the proposed ETFPOS-IDF outperformed all the schemes introduced by earlier studies in examination question classification and achieved 0.749 in accuracy and 0.746 in F1 score. The finding of this study demonstrates that distinguishing between different verb types is significant in reducing the misclassification of examination questions. This research contributed by introducing a novel term weighting scheme in classifying examination questions based on BT. Future work may involve identifying the optimal weight for both types of verbs, evaluating the proposed scheme with a larger dataset, and comparing the performance with deep learning.

INDEX TERMS Bloom's Taxonomy (BT), BT verbs, examination question classification, TF-IDF, term weighting.

I. INTRODUCTION

Educational data mining extracts valuable information from the raw data coming from educational systems [1]. Recent years have witnessed the rise of educational data mining applications. These applications involve predicting student performance [2] and motivation [3], student modeling [4], student behavior modeling [5], and many more. In addition, predicting the cognitive level of examination questions [6] is one of the educational data mining applications and the focus of this study.

Bloom's Taxonomy (BT) is a framework used in educational institutions to produce examination questions of various cognitive levels. Benjamin Bloom, an American educational psychologist, proposed this framework in 1956 [7]. The cognitive domain of BT consists of six levels, as shown in Fig. 1. These levels are ordered from low to high-order thinking, with Knowledge being the easiest and Evaluation being

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang.

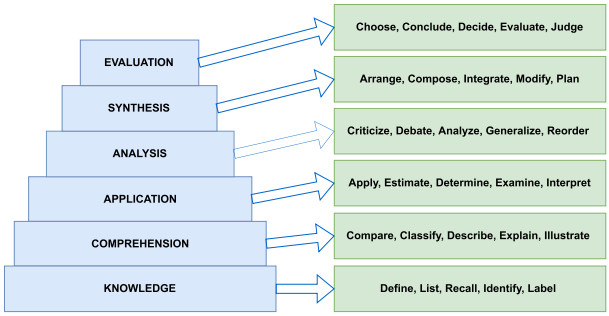

FIGURE 1. BT levels and some BT verbs corresponding to each level.

the most complex. Every examination question falls into one of these cognitive levels. However, using this framework to label the questions manually is time-consuming [8], [9], [10]. So, many past studies worked on examination question classification to automate the process using machine learning (ML) [8], [11], [12], Deep learning [10], and rule-based [13], [14] classification techniques. However, rule-based classification is not practical since new rules must be established whenever new data is added. In the classification of examination questions, there is no publicly accessible large-volume dataset, and labeling the questions needed a thorough understanding of the BT cognitive domain. So, building a large dataset is challenging and requires considerable time. Hence, this work focused on ML-based classification by introducing a novel term weighting scheme considering deep learning models often require a large amount of data to perform effectively.

The past studies on ML-based examination question classification tried to decrease the classification error by working on feature selection, feature extraction, and term weighting. Still, there are chances to increase the classification accuracy by working on term weighting since only a few studies worked on term weighting in examination question classification using BT. Term weighting is an approach to assigning numerical weight to the terms present in a document. Term weighting is inevitable since the text cannot be fed directly into the machine learning classifier to train and test the model. These numerical values represent the significance of those terms in the classification; the higher the value, the higher its importance is.

An examination question may contain BT verbs, supporting verbs, nouns, adjectives, adverbs, and many more. These parts of speech (POS) are not similarly significant while labeling or classifying the questions according to the BT. While categorizing the questions, verbs are the most important among the POS [11]. However, two types of verbs, such as BT verbs and supporting verbs, may be present in the question. BT verbs are more significant in determining the cognitive level of questions than supporting verbs since there is a strong link between BT verbs and BT levels. Fig. 1 shows some BT verbs with their corresponding cognitive levels. The past studies [11], [15], that emphasized verbs while weighting the terms did not distinguish between these two categories of verbs. However, distinguishing between these two categories of verbs may increase classification accuracy. So, this study proposed a new weighting scheme by distinguishing between the two categories of verbs mentioned earlier.

TABLE 1. Past studies of examination question classification on feature selection and feature set extraction.

Work	Year	Feature Selection	Feature Set Extraction
[16]	2010	✓	
[17]	2012	✓	
[18]	2014		✓
[19]	2015	✓	
[20]	2016		✓
[9]	2019		✓

II. RELATED WORK

This section divided the past studies of examination question classification into works on feature selection, feature set extraction, and term weighting. Table 1 shows the past studies of examination question classification on feature selection and feature set extraction, whereas term weighting is in Table 2.

A. WORK ON FEATURE SELECTION AND FEATURE SET EXTRACTION

To reduce the feature space for the Artificial Neural Network (ANN), [16] investigated a few feature selection methods, such as document frequency (DF) and category frequency-document frequency (CF-DF). The experiment result showed DF as a suitable feature reduction method in classifying examination questions. In contrast, this study found CF-DF inappropriate since it attempts to exclude verbs that could exist at more than one cognitive level. Reference [17] investigated the effectiveness of term frequency (TF) as a feature selection method and found that greater than or equal to two is the most optimal value for the TF. According to [18], most past studies focused on the bag-of-words (BOW) and syntactic features. So, this work introduced several features: keywords of the questions, headword or BT verb, syntactic and semantic. However, this research did not investigate the effects of these features in classifying examination questions.

Reference [19] tested multiple feature selection methods: Chi-Square, Mutual Information, and Odd Ratio in combination with various classifiers. These classifiers are Support Vector Machine (SVM), Naïve Bayes (NB), and k-Nearest Neighbour (KNN). The experiment results of this study showed the superiority of Mutual Information with a weighted feature size of 250. Reference [20] investigated whether individual or the combination of linguistically motivated features can increase classification accuracy or not. These features are Unigrams, Bigrams, Trigrams, POS Bigrams, POS Trigrams, and Word/POS Pairs. The experiment results of this work showed that Unigrams outperformed all other features in the single feature set experiment. The combination of Unigrams and Bigrams achieved the highest

TABLE 2. Past studies of examination question classification on term weighting.

Work	Year	Scheme	Approach
[13]	2012	Category weighting	Rule-based
[21]	2013	TF-IDF	ML-based
[22]	2015	Category weighting	Rule-based
[20]	2016	TF-IDF	ML-based
[23]	2016	Category weighting	Rule-based
[24]	2017	Binary	ML-based
[15]	2018	ETF-IDF	ML-based
[11]	2020	TFPOS-IDF	ML-based

result with the logistic regression (LR) classifier in the combination of feature sets test. Reference [9] extracted and tested different forms of TF-IDF: Words TF-IDF, N-Gram TF-IDF, and Characters TF-IDF with the NB classifier. The outcome of this study showed that the N-gram TF-IDF outperformed other forms of TF-IDF.

B. WORK ON TERM WEIGHTING

To solve the overlapping keyword problem of BT, [13] introduced category weighting for the conflicting category in rule-based classification. In this work, subject matter experts assigned the weights based on the question category. Another rule-based study [22] introduced category weighting for examination questions and assigned the weights based on the highest path and lemma similarities. Another rule-based study [23] used category weighting to classify the questions. However, they used wordnet and cosine similarity to assign weight to the question category.

Some past studies [20], [21] of machine learning-based examination question classification utilized the standard TF-IDF as a term weighting scheme. Reference [24] mentioned that TF and TF-IDF work well in situations where the words are repetitive. However, which is not the case in examination question classification since questions contain fewer terms. So, [24] applied the binary term weighting instead of TF and TF-IDF. Not all the words in the questions are equally significant while classifying the examination question. So, [15] introduced enhanced term frequency-inverse document frequency (ETF-IDF) and assigned a higher weight to the verbs present in the question compared to the other POS. Nouns and adjectives received higher weight than others POS. The outcome of this study showed that ETF-IDF outperformed traditional TF-IDF. In a later study [11], the same authors came up with a new scheme called TFPOS-IDF. TFPOS-IDF assigned the highest weight to the verbs, followed by the nouns and adjectives. The experiment result showed that TFPOS-IDF outperformed TF-IDF. However, this work performed no comparison between the ETF-IDF and TFPOS-IDF.

C. RESEARCH GAP IN TERM WEIGHTING

From the above discussion, it is observable that the schemes ETF-IDF and TFPOS-IDF assigned the highest weight to the verbs, followed by the nouns and adjectives. However, there could be more than one type of verb in questions: supporting and BT. The differences between these verb types are explained below with the help of an examination question.

Sample Question: "Suggest any (2) efforts that the organization may perform to discourage unethical behavior."

In the above sample question, the word 'suggest' at the beginning of the question is a BT verb. Two more verbs are also in the question: 'perform' and 'discourage.' However, these verbs are the supporting verbs. The schemes ETF-IDF and TFPOS-IDF did not distinguish between the BT verb and supporting verb. These schemes assigned equal weight to all the verbs, whether BT verb or supporting verb. However, discrimination between the different types of verbs may increase the classification accuracy since BT verbs present in the questions have a substantial impact in determining the cognitive levels of the questions than the supporting verbs. Nevertheless, no past studies addressed this issue during term weighting. Therefore, this study aims to introduce a novel term weighting scheme by identifying the BT verbs from the questions to assign a higher weight to the BT verbs than the supporting verbs.

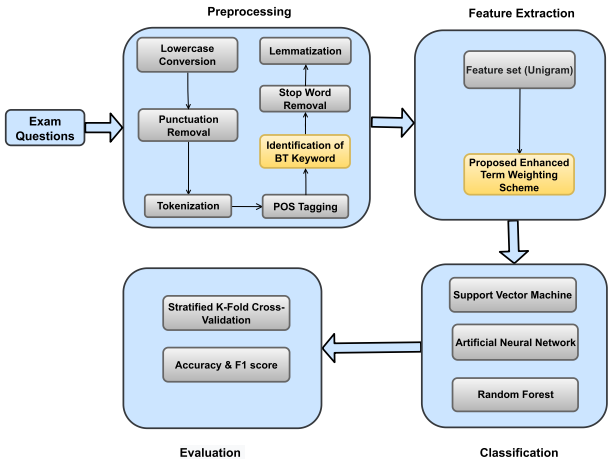

FIGURE 2. All the steps of the proposed examination question classification model.

III. METHODOLOGY

Fig. 2 illustrates all the steps involved in the proposed examination question classification model. The steps involved in the proposed solution are preprocessing the examination questions, feature extraction, classification, and model evaluation. Feature extraction involved creating the feature set and calculating the weights by applying the proposed term weighting scheme.

A. DATASET

This research utilized three datasets from earlier studies to train and test the examination question classification model. Pedagogy experts have already labeled these datasets according to the cognitive level of BT. The first dataset was introduced by [24] and consisted of 181 questions. Reference [24]

TABLE 3. Positions of BT verbs in the questions.

Position of BT Verb	Question	BT Verb	Supporting verb
The first word of a question	"Suggest any (2) efforts that the organization may perform to discourage unethical behavior."	Suggest	Perform, Discourage
The second word of a question, followed by an adverb	"Briefly explain any TWO (2) observations of information technology trend using Moore's Law."	Explain	Using
The second word of a question, followed by an adverb, and after the conjunction 'AND'	"Critically analyze the image above and discuss the strategic decisions associated with the appeals of the advertisement."	Analyze, Discuss	Associated
The first word of a question and, after the conjunction, 'AND'	"Argue the case for conducting experimental research involving humans and propose guidelines to ensure that the dignity and welfare of the subjects are maintained."	Argue, Propose	Conducting, Involving, Ensure, Maintained
Joined by the conjunction 'AND'	"Propose and justify ONE (1) international market-entry strategy that you may consider to market your products internationally."	Propose, Justify	Consider, Market

also introduced the second dataset comprising 415 questions from multiple fields such as Computing, Social Science, Business, and many more. The third dataset used in this study was introduced by [17] and consisted of 600 questions. However, many questions of the third dataset do not contain BT verbs. So, filtering was applied to remove the questions which do not have at least one BT verb. After discarding, 387 questions remained in the dataset. Though few studies [9], [25] categorized the BT levels into low and high order, this study used the six levels of the cognitive domain as class labels in the target variable for classification.

B. PREPROCESSING

Preprocessing the text data involves many steps. These steps include the elimination of punctuation and stopwords, tokenization, stemming or lemmatization, POS, and many more. Many past studies [19], [20], [22] used these methods to preprocess the examination questions. Besides these, the BT verbs in every question need to identify since a different weight needs to assign to the BT verbs than the supporting verbs in the proposed term weighting scheme.

In this study, at first, converted the questions into lowercase. The punctuations present in the question were removed and tokenized the questions. After that, the pos tagging was applied to the terms using the Stanford tagger (version 4.2.0) [26] by following past studies [11], [15], [27]. The BT verbs need to identify to use later in the proposed scheme of this study. So, all the BT verbs were determined by their position in the questions, as shown in Table 3. The detailed process of identifying the BT verbs is discussed in section III-C2.a. After identifying the BT verbs, stop words were removed, followed by the lemmatization of the terms. The stop word list of the NLTK (version 3.6.1) [28] was used in this study, whereas the WordNetLemmatizer for the lemmatization.

C. FEATURE EXTRACTION

1) FEATURE SET

Before calculating the term weighting values of all the terms present in a question, there should be a feature set. This study used the unigram to obtain a feature set containing all the unique terms present in the dataset. Past studies of examination question classification, especially those studies [11], [15] that worked on term weighting, also used unigram.

2) TERM WEIGHTING

This study implemented three schemes proposed by past studies to compare the performance of the proposed scheme with past schemes of examination question classification. These schemes are TF-IDF, ETF-IDF [15], and TFPOS-IDF [11]. Among these three, ETF-IDF and TFPOS-IDF are the two latest schemes proposed in examination question classification. The standard TF-IDF has many variations. This study used the most optimal variant of TF-IDF identified by [29].

a: PROPOSED TERM WEIGHTING SCHEME ETFPOS-IDF

The proposed scheme ETFPOS-IDF is the enhanced version of the TFPOS-IDF proposed by [11]. The TFPOS-IDF discriminates between the POS and assigns a higher weight to the verbs. However, TFPOS-IDF does not differentiate between the different types of verbs, such as BT verbs and supporting verbs. The proposed scheme of this study discriminated between the types of verbs and assigned higher weights to the BT verbs than the supporting verbs. The ETFPOS-IDF is discussed in (1) to (3).

$$Ew_{pos}(t) = \left\{ \begin{array}{c} w1, \text{ if } t \text{ is BT } Verb \\ w2, \text{ if } t \text{ is Supporting } Verb \\ w3, \text{ if } t \text{ is } Noun \text{ or } Adjective \\ w4, \text{ otherwise} \end{array} \right\} \quad (1)$$

where w1 = 5, w2 = 3, w3 = 2, and w4 = 1. So, in (1), the BT verbs were assigned weight value 5, where 3 to the supporting verbs. The BT verbs were identified by their position in the questions, as shown in Table 3 earlier.

We analyzed the questions from all three datasets to identify the BT verbs from the questions. From the questions, we found some patterns to identify the BT verbs, as demonstrated in Table 3. There was no other way to identify the BT verbs except by analyzing the positions of the verbs in the questions. Therefore, we identified all the BT verbs in the

Algorithm 1 Process in Identifying BT Verbs

```
q: A Question
d: BT Verbs Database
function Identify(q, d)
    sentences ← split question
    newlist ← []
    for sentence in sentences do
        words ← split sentence
        x ← first word of the sentence
        if x is in d then
            newlist.insert((x, "BT"))
        else
            newlist.insert((x, "non-BT"))
        end if
        for word in remaining words do
            if word is in d then
                y ← previous word
                if y == "and" then
                    newlist.insert((word, "BT"))
                else if y is (Adverb & endswith "ly") then
                    newlist.insert((word, "BT"))
                else
                    newlist.insert((word, "non-BT"))
                end if
            else
                newlist.insert((word, "non-BT"))
            end if
        end for
    end for
    return newlist
end function
```

questions using the BT verb's position. This process was very tedious and time-consuming.

The implementation to identify the BT verbs are illustrated in Algorithm 1. At first, the questions were split into sentences based on commas, semicolons, and full stops. After that, every sentence was split into words. The first word of every sentence was searched in the BT verbs database to identify whether it was a BT verb or not. We have collected the BT verbs database from past research [24]. The first word was added to the list with the label BT if it was a BT verb. If not, it was then labeled as non-BT. For the remaining words, every word was labeled as BT if the word present in the BT verbs database and the previous word of that word is 'and' or an adverb ends with 'ly'; otherwise labeled as non-BT. After that, we manually examined each dataset's output to ensure that the BT verbs had been appropriately identified.

Finally, we compared the aforementioned algorithm's output with the POS-tagged preprocessed questions from the preprocessing phase described in section III-B to replace the label of non-BT words with their POS. So, in this process, the BT verbs remained with the label BT; however, the label of non-BT words was replaced with the POS of that word. After that, the stop words removal and lemmatization processes were performed, as mentioned earlier in the preprocessing stage.

The calculated $Ew_{pos}(t)$ from (1) was used to calculate the $ETFPOS(t, q)$, as shown in (2).

$$ETFPOS(t, q) = \frac{C\ (t, q) \times\ Ew_{pos}(t)}{\sum_i C(t_i, q) \times Ew_{pos}(t_i)} \tag{2}$$

where $C\ (t, q)$ represents the frequency of t in question q and $\sum_i C(t_i,q)$ is the total number of terms in question q.

Finally, $ETFPOS - IDF(t, q)$ was calculated using (3).

$$ETFPOS - IDF(t, q) = ETFPOS(t, q) \cdot IDF(t) \tag{3}$$

The $ETFPOS - IDF(t, q)$ is the multiplication of $ETFPOS(t, q)$ and $IDF(t)$, as shown in (3).

The normalization technique prevents the numerical complexity of calculation during the model training process, as stated by [11]. This study normalized the weighting values of the proposed scheme ETFPOS-IDF using the L2 normalization technique. As a result, all the weighting values converted between 0 and 1. The TFPOS-IDF has also been normalized by following the [11]. The normalized term weighting values were obtained using (4).

$$Normalized\ ETFPOS - IDF(t, q) = \frac{ETFPOS - IDF(t, q)}{\sqrt{\sum ETFPOS - IDF(t, q)^2}} \tag{4}$$

In (4), $ETFPOS - IDF(t, q)$ is the term weighting value obtained for t in question q.

D. CLASSIFICATION AND EVALUATION

This study used three famous machine learning classifiers: SVM, Random Forest (RF), and ANN. The extensively used [11], [15], [24] Python module, Scikit-learn (version 1.0.1) [30], was used in this study to train and test the classifier.

SVM: was introduced by [31] in machine learning to solve classification problems. SVM has been widely used in text and examination question classification [19], [21], [32]. The past studies [15], [27] of examination question classification used the linear kernel of SVM, also known for higher accuracy in text classification [33]. Hence, this study used the linear kernel of SVM with the default settings of Scikit-learn.

RF: RF is one of the most effective classifiers for text classification [34], introduced by Leo Breiman [35]. Several past studies [36], [37], [38] used RF for text classification purposes. RF is an ensemble classifier based on decision trees, and it uses the majority voting technique to determine the final predicted class [34]. One of the advantages of RF is that it can handle the overfitting issue [39], which was an issue in the decision tree classifier. This study used the Scikit-learn implementation of RF with the default settings and the random state as 42 for the reproducible results.

ANN: This classifier, also known as Multilayer Perceptron, was used in many past studies [40], [41] of text classification. ANN consists of the input, hidden, and output layers. There

TABLE 4. Experiment results of SVM.

Term weighting/ Dataset	TF-IDF		ETF-IDF (2018)		TFPOS-IDF (2020)		ETFPOS-IDF (Proposed)	
	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
Dataset 1	0.698	0.684	0.713	0.707	0.733	0.731	**0.750**	**0.748**
Dataset 2	0.629	0.616	0.684	0.680	0.689	0.680	**0.700**	**0.696**
Dataset 3	0.733	0.729	0.798	0.795	0.807	0.804	**0.843**	**0.840**

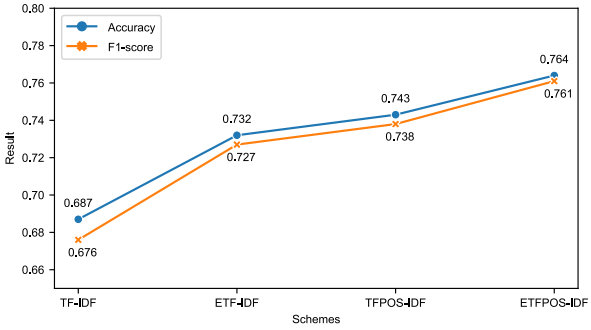

FIGURE 3. Average result of each scheme with SVM.

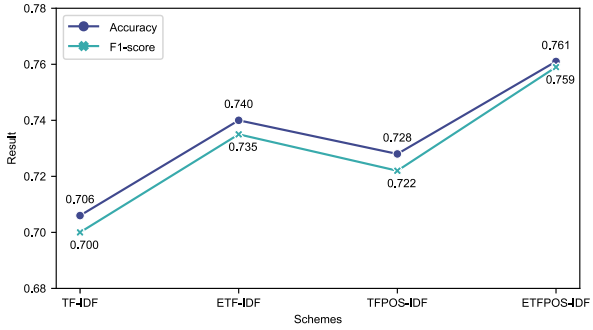

FIGURE 4. Average result of each scheme with ANN.

could be more than one hidden layer in ANN. However, this study used the default setting for the number of hidden layers and neurons of the ANN classifier available in Scikit learn. As a random state for ANN, zero was used to achieve reproducible results. Besides this, 'lbfgs' was used as a solver since it converges faster with the small dataset, according to Scikit-learn [30] documentation of ANN.

Evaluation Metrics and Cross-validation: This study used accuracy and F1 score as evaluation metrics to measure the performance of the proposed model. Many past studies [17], [21], [32] of examination question classification used these metrics. To split the dataset into training and test set, we used the stratified k-fold cross-validation technique with the random state as 0. Stratified cross-validation ensures that each fold contains about the exact proportion of data points from each class label present in the dataset [42]. This study adopted the approach of [27] to use multiple k-values to achieve more reliable results. As k-values, a range from 3 to 10 was used in incremental order. The final value was determined by calculating the mean for each k-value and then the mean for all k-values.

IV. RESULTS AND DISCUSSION

A. EXPERIMENT RESULTS OF SVM

Table 4 presents the proposed scheme's results and the results of term weighting schemes used in the past studies of examination question classification based on the BT with the SVM classifier. The table shows that in every dataset, the proposed term weighting scheme ETFPOS-IDF outperformed the traditional TF-IDF, ETF-IDF, and TFPOS-IDF. In all the datasets, TFPOS-IDF performed closely to the proposed ETFPOS-IDF. If we compare the average performance of each scheme with the SVM classifier, we can see from Fig. 3 that the ETFPOS-IDF outperformed all and achieved 0.764 in accuracy and 0.761 in F1 score. Among the past schemes, TFPOS-IDF performed closely to the ETFPOS-IDF, and the difference is 2.1 % and 2.3 % in accuracy and F1 score, respectively. The traditional TF-IDF performed the least satisfactorily among the four schemes. These results show that the proposed ETFPOS- IDF improves the classification accuracy of examination question classification with SVM.

B. EXPERIMENT RESULTS OF ANN

Table 5 demonstrates the experiment results of the proposed ETFPOS-IDF along with other schemes with the ANN classifier. From the results, the proposed scheme ETFPOS-IDF outperformed all the schemes in all the datasets used in this study. However, in Dataset 2, the difference in performance between the ETFPOS-IDF and ETF-IDF is identical, approximately 1.5 % in both metrics. With the ANN classifier, if we compare the average performance of each dataset, we can observe from Fig. 4 that the proposed term weighting scheme ETFPOS-IDF outperformed all and achieved an average accuracy of 0.761 and an average F1 score of 0.759. Like the SVM classifier, the proposed scheme ETFPOS-IDF shows improvement in classifying examination questions based on BT with ANN.

C. EXPERIMENT RESULTS OF RF

Table 6 illustrates the experiment results of the RF for all the datasets used in this research. The results show that in Datasets 1 and 3, the proposed ETFPOS-IDF outperformed all the other schemes. However, in Dataset 2, TFPOS-IDF surpassed all, including the proposed ETFPOS-IDF. In Dataset 1, the difference in performance between TFPOS-IDF and ETFPOS-IDF is minimal, 0.5% in both

TABLE 5. Experiment results of ANN.

Term weighting/ Dataset	TF-IDF		ETF-IDF (2018)		TFPOS-IDF (2020)		ETFPOS-IDF (Proposed)	
	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
Dataset 1	0.697	0.686	0.714	0.702	0.707	0.695	**0.736**	**0.732**
Dataset 2	0.670	0.667	0.701	0.698	0.696	0.693	**0.715**	**0.714**
Dataset 3	0.751	0.747	0.806	0.804	0.780	0.778	**0.833**	**0.832**

TABLE 6. Experiment results of RF.

Term weighting/ Dataset	TF-IDF		ETF-IDF (2018)		TFPOS-IDF (2020)		ETFPOS-IDF (Proposed)	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Dataset 1	0.694	0.692	0.688	0.684	0.702	0.699	**0.707**	**0.704**
Dataset 2	0.661	0.649	0.661	0.649	**0.689**	**0.680**	0.672	0.662
Dataset 3	0.741	0.736	0.740	0.735	0.768	0.764	**0.791**	**0.788**

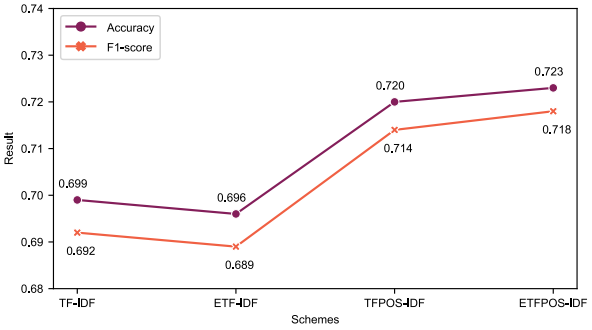

FIGURE 5. Average result of each scheme with RF.

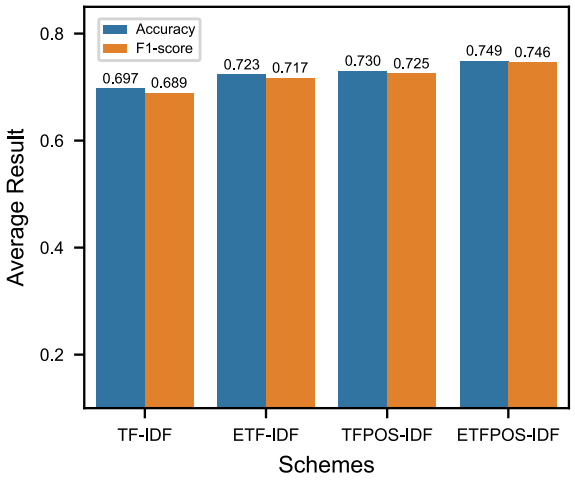

FIGURE 6. Average performance of each of the schemes.

metrics. However, in Dataset 2, there is a considerable difference between these two schemes in performance, 2.3% in accuracy and 2.4% in F1 score. From Fig. 5, we can see that overall, the proposed ETFPOS-IDF outperformed all the schemes. However, the performance of TFPOS-IDF is nearly identical to ETFPOS-IDF.

D. SUMMARY

Fig. 6 illustrates the summary of the results. The result is obtained by averaging the results of all classifiers and datasets used in this study. From Fig. 6, it is observable that the proposed scheme ETFPOS-IDF outperformed all other schemes. The ETFPOS-IDF outperformed the closest performed TFPOS-IDF by approximately 1.9% in accuracy and 2.1% in F1 score. If we compare ETF-IDF and ETFPOS-IDF, ETFPOS-IDF outperformed ETF-IDF by around 2.6% and 2.9% in accuracy and F1 score, respectively. The difference is even higher with the TF-IDF, approximately 5.2% and 5.7% in accuracy and F1 score, respectively.

E. DISCUSSION

From the experiment result, we have found that the proposed scheme ETFPOS-IDF outperformed the other schemes of examination question classification proposed by previous studies, such as ETF-IDF, TFPOS-IDF, and standard

TABLE 7. Term weighting values of proposed ETFPOS-IDF and other schemes for a question.

Terms/ Scheme	list (BT verb)	step	involve (Supporting verb)	titration
TF-IDF	0.39348	0.499809	0.545603	0.545603
ETF-IDF	0.541701	0.253646	0.751998	0.276988
TFPOS-IDF	0.4882	0.372075	0.676942	0.406165
ETFPOS-IDF	0.665491	0.338131	0.553666	0.369111

TF-IDF. In ETF-IDF and TFPOS-IDF, higher weights were assigned to the verbs compared to the other POS, as verbs are more significant than any other POS while determining the cognitive levels of examination questions. However, there was no discrimination between the supporting verbs and BT verbs while weighting the terms in ETF-IDF and TFPOS-IDF. Table 7 presents the weighting values of all the schemes for a question taken randomly from Dataset 3. The table shows that the difference between the BT and supporting verbs is higher in the proposed scheme ETFPOS-IDF compared to

TF-IDF, ETF-IDF, and TFPOS-IDF. The reason is that in ETFPOS-IDF, the BT verbs received higher weights than the supporting verbs. The better performance of ETFPOS-IDF could result from discriminating between the type of verbs while weighting the terms.

V. CONCLUSION

This study proposed a new term weighting scheme ETFPOS-IDF in examination question classification based on BT. The proposed ETFPOS-IDF distinguished the type of verbs present in the questions and assigned a higher weight to the BT verbs than the supporting verbs. This study used three datasets and three classifiers to investigate the effectiveness of the proposed scheme. The results of the proposed ETFPOS-IDF were compared with schemes devised by earlier studies. These schemes are traditional TF-IDF, ETF-IDF, and TFPOS-IDF. Accuracy and F1 score, two widely used classification evaluation metrics, were used in this study to evaluate the results. The results of the classifiers showed that the proposed scheme outperformed all other schemes with the SVM, ANN, and RF classifiers. This outcome indicated that distinguishing the type of verbs while weighting the terms increases the accuracy significantly in classifying examination questions. The later study may utilize a larger dataset to evaluate the stability of the proposed scheme. Future studies can also identify the optimal weight difference between the types of verbs and compare the performance of the proposed scheme with the deep learning approach.

REFERENCES

[1] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010, doi: 10.1109/TSMCC.2010.2053532.

[2] Z. Xu, H. Yuan, and Q. Liu, "Student performance prediction based on blended learning," *IEEE Trans. Educ.*, vol. 64, no. 1, pp. 66–73, Feb. 2021, doi: 10.1109/TE.2020.3008751.

[3] M. Munoz-Organero, P. J. Munoz-Merino, and C. D. Kloos, "Student behavior and interaction patterns with an LMS as motivation predictors in E-learning settings," *IEEE Trans. Educ.*, vol. 53, no. 3, pp. 463–470, Aug. 2010, doi: 10.1109/TE.2009.2027433.

[4] F. Alshareef, H. Alhakami, T. Alsubait, and A. Baz, "Educational data mining applications and techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 4, pp. 729–734, 2020, doi: 10.14569/IJACSA.2020.0110494.

[5] S. L. Prabha and A. R. M. Shanavas, "Educational data mining applications," *Oper. Res. Appl., Int. J.*, vol. 1, no. 1, pp. 23–29, 2014.

[6] S. Hasnah, S. Fattah, R. S. Sulong, and M. Mamat, "Mining exam question based on Bloom's taxonomy," in *Proc. Knowl. Manag. Int. Conf.*, 2008, pp. 424–427.

[7] B. S. Bloom, *Taxonomy of Educational Objectives: The Classification of Educational Goals*. New York, NY, USA: David McKay Company, 1956.

[8] S. F. Kusuma, D. Siahaan, and U. L. Yuhana, "Automatic Indonesia's questions classification based on Bloom's taxonomy using natural language processing a preliminary study," in *Proc. Int. Conf. Inf. Technol. Syst. Innov.* 2016, pp. 1–6, doi: 10.1109/ICITSI.2015.7437696.

[9] A. Aninditya, M. A. Hasibuan, and E. Sutoyo, "Text mining approach using TF-IDF and naive Bayes for classification of exam questions based on cognitive level of Bloom's taxonomy," in *Proc. IEEE Int. Conf. Internet Things Intell. Syst. (IoTaIS)*, Nov. 2019, pp. 112–117, doi: 10.1109/IOTAIS47347.2019.8980428.

[10] S. Shaikh, S. M. Daudpotta, and A. S. Imran, "Bloom's learning Outcomes' automatic classification using LSTM and pretrained word embeddings," *IEEE Access*, vol. 9, pp. 117887–117909, 2021, doi: 10.1109/ACCESS.2021.3106443.

[11] M. Mohammed and N. Omar, "Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec," *PLoS ONE*, vol. 15, no. 3, pp. 1–21, Mar. 2020, doi: 10.1371/JOURNAL.PONE.0230442.

[12] K. Osadi, M. Fernando, and W. Welgama, "Ensemble classifier based approach for classification of examination questions into Bloom's taxonomy cognitive levels," *Int. J. Comput. Appl.*, vol. 162, no. 4, pp. 1–6, Mar. 2017.

[13] N. Omar, S. S. Haris, R. Hassan, H. Arshad, M. Rahmat, N. F. A. Zainal, and R. Zulkifli., "Automated analysis of exam questions according to Bloom's taxonomy," *Proc. Soc. Behav. Sci.*, vol. 59, no. 1956, pp. 297–303, 2012, doi: 10.1016/J.SBSPRO.2012.09.278.

[14] W.-C. Chang and M.-S. Chung, "Automatic applying Bloom's taxonomy to classify and analysis the cognition level of English question items," in *Proc. Joint Conf. Pervasive Comput. (JCPC)*, Dec. 2009, pp. 727–734, doi: 10.1109/JCPC.2009.5420087.

[15] M. Mohammed and N. Omar, "Question classification based on Bloom's Taxonomy using enhanced TF-IDF," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 8, nos. 2–4, pp. 1679–1685, 2018, doi: 10.18517/IJASEIT.8.4-2.6835.

[16] N. Yusof and C. J. Hui, "Determination of Bloom's cognitive level of question items using artificial neural network," in *Proc. 10th Int. Conf. Intell. Syst. Design Appl.*, 2010, pp. 866–870, doi: 10.1109/ISDA.2010.5687152.

[17] A. A. Yahya, Z. Toukal, and A. Osman, "Bloom's taxonomy–based classification for item bank questions using support vector machines," in *Modern Advances in Intelligent Systems and Tools*, vol. 431, 2012, pp. 135–140.

[18] A. Sangodiah, R. Ahmad, and W. F. W. Ahmad, "A review in feature extraction approach in question classification using support vector machine," in *Proc. IEEE Int. Conf. Control Syst., Comput. Eng. (ICCSCE)*, Nov. 2014, pp. 536–541, doi: 10.1109/ICCSCE.2014.7072776.

[19] D. A. Abduljabbar and N. Omar, "Exam questions classification based on Bloom's taxonomy cognitive level using classifiers combination," *J. Theor. Appl. Inf. Technol.*, vol. 78, no. 3, pp. 447–455, Aug. 2015.

[20] A. Osman and A. A. Yahya, "Classifications of exam questions using linguistically-motivated features: A case study based on Bloom's taxonomy," in *Proc. 6th Int. Arab Conf. Quality Assurance Higher Educ.* 2016, vol. 2016.

[21] A. A. Yahya, A. Osman, A. Taleb, and A. A. Alattab, "Analyzing the cognitive level of classroom questions using machine learning techniques," *Proc. Social Behav. Sci.*, vol. 97, pp. 587–595, Nov. 2013, doi: 10.1016/J.SBSPRO.2013.10.277.

[22] K. Jayakodi, M. Bandara, and I. Perera, "An automatic classifier for exam questions in engineering: A process for Bloom's taxonomy," in *Proc. IEEE Int. Conf. Teaching, Assessment, Learn. Eng. (TALE)*, Dec. 2015, pp. 195–202, doi: 10.1109/TALE.2015.7386043.

[23] K. Jayakodi, M. Bandara, I. Perera, and D. Meedeniya, "WordNet and cosine similarity based classifier of exam questions using Bloom's taxonomy," *Int. J. Emerg. Technol. Learn.*, vol. 11, no. 4, pp. 142–149, Apr. 2016, doi: 10.3991/IJET.V11I04.5654.

[24] A. Sangodiah, R. Ahmad, and W. F. W. Ahmad, "Taxonomy based features in question classification using support vector machine," *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 12, pp. 2814–2823, 2017.

[25] A. J. Swart, "Evaluation of final examination papers in engineering: A case study using Bloom's taxonomy," *IEEE Trans. Educ.*, vol. 53, no. 2, pp. 257–264, May 2010, doi: 10.1109/TE.2009.2014221.

[26] *The Stanford Natural Language Processing Group*. Accessed: Aug. 15, 2022. [Online]. Available: https://nlp.stanford.edu/software/tagger.shtml

[27] A. Sangodiah, Y. T. Fui, L. E. Heng, N. A. Jalil, R. K. Ayyasamy, and K. H. Meian, "A comparative analysis on term weighting in exam question classification," in *Proc. 5th Int. Symp. Multidisciplinary Stud. Innov. Technol. (ISMSIT)*, Oct. 2021, pp. 199–206, doi: 10.1109/ISMSIT52890.2021.9604639.

[28] S. Bird, E. Klein, and E. Loper, *Natural Language Processing With Python: Analyzing Text With the Natural Language Toolkit*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2009.

[29] A. Sangodiah, T. J. San, Y. T. Fui, L. E. Heng, R. K. Ayyasamy, and N. A. Jalil, "Identifying optimal baseline variant of unsupervised term weighting in question classification based on Bloom taxonomy," *MENDEL*, vol. 28, no. 1, pp. 8–22, Jun. 2022.

[30] F. Pedregosa, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.

[31] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[32] A. A. Yahya and A. Osman, "Automatic classification of questions into Bloom's cognitive levels using support vector machines," in *Proc. Int. Arab Conf. Inf. Technol.*, 2011, pp. 1–6.
[33] M. Pota, M. Esposito, and G. De Pietro, "A forward-selection algorithm for SVM-based question classification in cognitive systems," *Smart Innov. Syst. Technol.*, vol. 55, pp. 587–598, Jan. 2016, doi: 10.1007/978-3-319-39345-2_52/COVER.
[34] A. Bouaziz, C. Dartigues-Pallez, C. da Costa Pereira, F. Precioso, and P. Lloret, "Short text classification using semantic random forest," in *Proc. Int. Conf. Data Warehousing Knowl. Discover*, 2014, pp. 288–299, doi: 10.1007/978-3-319-10160-6_26.
[35] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[36] X. Luo, "A new text classifier based on random forests," in *Proc. 2nd Int. Conf. Mater. Eng. Inf. Technol. Appl. (MEITA)*, 2017, pp. 290–293, doi: 10.2991/MEITA-16.2017.60.
[37] D. P. Kavadi, P. Ravikumar, and K. S. Rao, "A new supervised term weight measure for text classification," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 6, pp. 3115–3128, 2020.
[38] G. Domeniconi, G. Moro, R. Pasolini, and C. Sartori, "A study on term weighting for text categorization: A novel supervised variant of tf.idf," in *Proc. 4th Int. Conf. Data Manage. Technol. Appl.*, 2015, pp. 26–37, doi: 10.5220/0005511900260037.
[39] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random forest," in *Proc. Int. Conf. Inf. Comput. Appl.*, 2012, pp. 246–252, doi: 10.1007/978-3-642-34062-8_32.
[40] R. F. de Mello, L. J. Senger, and L. T. Yang, "Automatic text classification using an artificial neural network," *IFIP Adv. Inf. Commun. Technol.*, vol. 172, pp. 215–238, Jan. 2005, doi: 10.1007/0-387-24049-7_12/COVER.
[41] P. L. Prasanna and D. R. Rao, "Text classification using artificial neural networks," *Int. J. Eng. Technol.*, vol. 7, no. 1, pp. 603–606, 2018, doi: 10.14419/IJET.V7I1.1.10785.
[42] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of Database System*. Boston, MA, USA: Springer, 2009, pp. 532–538, doi: 10.1007/978-0-387-39940-9_565.

MOHAMMED OSMAN GANI was born in Chattogram, Bangladesh, in 1998. He received the B.Sc. degree in data science from Multimedia University, Malaysia, in 2021. He is currently pursuing the M.S. degree in computer science with Universiti Tunku Abdul Rahman, Malaysia.

RAMESH KUMAR AYYASAMY (Senior Member, IEEE) received the Ph.D. degree in information technology from Monash University, Australia, in 2013.

Starting from 2003 to 2008 and from 2013 to 2014, he worked as a Lecturer in Tamil Nadu, India, and Monash University, Malaysia, consecutively. Since 2015, he has been working as an Assistant Professor with the Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Malaysia. His research interests include artificial intelligence, big data analytics, cyberbullying, deep learning, machine learning, and text mining.

Dr. Ramesh serving as a reviewer for different journals and conferences and a member of editorial boards.

SAADAT M. ALHASHMI received the Ph.D. degree from Sheffield Hallam University, U.K. He is currently working as an Associate Professor with the Department of Information Systems, University of Sharjah, United Arab Emirates. He has supervised several Ph.D. students and published several papers in high-impact journals and conferences.

ANBUSELVAN SANGODIAH was born in Perak, Malaysia, in 1973. He received the bachelor's degree in science and the master's degree in information technology from the University of Putra Malaysia, in 1996 and 2000, respectively, and the Ph.D. degree in information technology from Universiti Teknologi PETRONAS.

He has been a Lecturer and Information Technology Trainer for the past 20 years. He is currently with the School of Computing, Faculty of Computing and Engineering, Quest International University. Prior to that, he worked in several renowned universities in Malaysia. He has been involved in research focusing on text classification, particularly in the education domain. He has published his research work in various international journals. His research interest includes computer science, particularly software development, database, and artificial intelligence.

Dr. Sangodiah is a Registered Professional Technologist with the Malaysia Board of Technologists (MBOT), Malaysia.

YONG TIEN FUI was born in Kuala Lumpur, Malaysia, in 1974. He received the M.A. degree in computing from the University of Aberdeen, Scotland, in 1998, and the M.B.A. degree in information technology from International Islamic University Malaysia, in 2002.

From 2000 to 2003, he was a Lecturer with Universiti Kuala Lumpur, Malaysia. From 2003 to 2006, he was a Lecturer at Multimedia University. From 2007 to 2008, he served two other institutions as an Information Technology Lecturer in Malaysia. He is currently a Lecturer of information systems with Universiti Tunku Abdul Rahman, Malaysia. His research interests include recommender systems, text mining, and natural language processing.

Mr. Yong is a Registered Professional Technologist with the Malaysia Board of Technologists (MBOT), Malaysia.